Theoretical Adequacy, Human Data and Classification Approaches in Modelling Word Properties, Word Relatedness and Word Classes

Sabine Schulte im Walde

Juni 2008

Habilitationsschrift gemäß Paragraph 1, Abschnitt 2 der gemeinsamen Habilitationsordnung der Philosophischen Fakultäten der Universität des Saarlandes vom 12. Juli 2000

Contents

Chapter 1

Introduction

The field of data-intensive lexical semantics is concerned with automatically deriving lexical semantic knowledge (i.e., the meaning of lexical units) from large-scale data resources.[1] The goals of the resulting descriptions are manifold: to define the semantics of words or multi-word expressions per se;[2] to define models of word properties and word relatedness; or to integrate the lexical information into applications in Natural Language Processing (NLP) that require semantic knowledge, such as anaphora resolution, question answering, summarisation, etc.

A major challenge within data-intensive lexical semantics is to bridge the gap between (1) the theoretical adequacy of the acquired semantic knowledge, (2) the potential to acquire the desired semantic knowledge from existing resources, and (3) the successful application of the semantic knowledge:

(1) **Theoretical standards:**

In general, there is pre-existing (and often, well-defined) knowledge about the linguistic foundations of the research questions under consideration. For example, there is an enormous amount of linguistic and psycholinguistic literature concerning verbs in general, and also concerning specific subclasses of verbs, such as particle verbs, light verbs, etc. For theoretically adequate models of word meaning, the models should take into account and integrate such pre-existing linguistic specifications.

(2) **Acquisition potential:**

Without any kind of semantic pre-processing, the acquisition of lexical semantic knowledge from corpus data is not trivial, and few resources –which, in addition, are typically available for some privileged languages only– are semantically annotated and provide semantic

[1] Strictly speaking, data-intensive lexical semantics refers to the automatic acquisition of lexical semantic knowledge from empirical data. Manual resources, however, are often exploited in addition to empirical data, thus we do not exclude them from large-scale data resources in this work.

[2] Henceforth, the terms *word* and *multi-word expression*, or only *word* will be used to refer to *lexical unit*.

information off-the-shelf (such as *FrameNet* (Fillmore *et al.*, 2003) and *PropBank* (Palmer *et al.*, 2005) for English). Thus, the automatic induction of lexical semantic knowledge either concentrates on semantic information that is trivial to induce from corpus data (such as adverbs as one hint towards the aktionsart of verbs), or benefits from a long-standing linguistic hypothesis which asserts a tight connection between the lexical meaning of a word and its distributional behaviour (Harris, 1968; Pinker, 1989; Levin, 1993), e.g., to induce semantic word relations based on lexico-syntactic co-occurrence patterns. In the former case, one is restricted to the subset of word properties that is directly available; in the latter case, one relies on a meaning-behaviour relationship that is not perfect, i.e, the heuristics are valid only to a certain extent, and thus introduce uncertainty into the semantic knowledge. Summarising, the acquisition potential of semantic lexical knowledge is typically a compromise between theoretically-based semantic properties and the availability of semantic information.

(3) **Successful model application:**

The automatic induction of semantic knowledge is often not an end task in itself, but is rather driven by the application scenario of the desired information. Thus, the theoretical adequacy of the semantic knowledge is of secondary importance; most relevant in such cases is the successful application of the model. The success of the model, however, is not necessarily correlated with the theoretical adequacy of the semantic information. Instead, the best models may rely on optimising technical rather than linguistic model parameters. An example of this is the best system in the second PASCAL Recognising Textual Entailment challenge (Bar-Haim *et al.*, 2006) by Hickl *et al.* (2006), which outperformed all other systems mainly because of increasing the training data, rather than outstanding semantic knowledge.

The degree to which an individual researcher is interested in taking the above three aspects into account varies, of course, depending on whether the focus of the research is on the theoretical adequacy of the semantic modelling, or on the success of the application.

The aim of this work is to bring together aspects (1), (2) and (3) of models in data-intensive lexical semantics –theoretical adequacy, the acquisition potential, and appropriate modelling–, with respect to semantic word classes. The focus thereby is on word properties and on definitions of relatedness between words that are crucial within the automatic acquisition of semantic classes, addressing characteristics of semantic classes as well as classification approaches that support the automatic acquisition of such classes. As a major source for the theoretical adequacy of the semantic knowledge, we exploit human data (associations and judgements), cf. Chapter 2, and apply them to semantic classification itself, and to word sense disambiguation of German nouns. In Chapter 3, we explore empirical word properties for specific instantiations of semantic classes, German particle verbs and Catalan adjectives. The final Chapter 4 presents classification models that address the polysemy of lexical units: multi-label classification, an ensemble classifier, and an Expectation-Maximisation (EM) classifier incorporating the Minimum Description Length (MDL) principle.

The contributions of this work can be summarised as follows.

(1) **Human data and the acquisition of semantic classes:**

Starting out with the questions why there are so many classifications with the same target objects, why and how they differ, and whether any of them is optimal, I suggest human data as an instrument (a) to identify and evaluate the semantic appropriateness of features within semantic classification, and (b) to assess the results of an automatic semantic classification process.

(2) **Feature exploration for specific semantic classifications:**

The theoretical adequacy of features within models of semantic relatedness and classification is addressed for two specific semantic classifications, German particle verbs and Catalan adjectives. Based on word properties derived from the theoretical literature, the correspondence between what theory predicts and the effect of the respective feature choice on the semantic models is explored.

(3) **Approaches to modelling polysemy in semantic classification:**

Addressing the lack of semantic classification models that so far have explicitly included polysemy into the automatic procedure, two approaches to modelling polysemy in semantic classes are suggested.

The remainder of the introduction will motivate and describe our work in some more detail. In parallel to the main chapters, Section 1.1 focuses on human data with respect to semantic class acquisition, Section 1.2 on the feature exploration[3] for specific instantiations of semantic classes, and Section 1.3 on classification approaches to modelling polysemy in semantic classes.

1.1 Human Data and the Acquisition of Semantic Classes

In recent years, the (computational) linguistics community has developed an impressive number of *semantic classifications*, i.e., classifications that generalise over their objects according to the objects' semantic properties. These classifications cover all major parts-of-speech, and manual as well as automatic definitions.

Major frameworks of manual classifications are *WordNet* (Miller, 1990; Fellbaum, 1998), a lexical semantic taxonomy that organises English nouns, verbs, adjectives and adverbs into classes of synonyms, and *FrameNet* (Baker *et al.*, 1998; Fillmore *et al.*, 2003), a database that is based on Fillmore's frame semantics (Fillmore, 1982) and describes the background and situational knowledge needed for understanding a word or expression. Both classifications started out for English and have subsequently been transferred to further languages. In addition, there is a large

[3]While we refer to semantic characteristics of lexical units from a theoretical point of view by the term *property*, we refer to resource-derived characteristics by the term *feature*, which is more common in the data-intensive literature.

number of manual classifications that have been developed for one specific language only, such as the Levin verb classes (Levin, 1993) and *VerbNet* (Kipper Schuler, 2006), a semi-automatic extension of the Levin classes, or the process-based classification of German verbs by Ballmer and Brennenstuhl (1986), to name just a few examples of verb classifications.

Concerning the automatic acquisition of semantic classes, we find approaches to noun classification such as Hindle (1990), approaches to verb classification such as Schulte im Walde (2000); Merlo and Stevenson (2001); Korhonen *et al.* (2003); Schulte im Walde (2006b); Joanis *et al.* (2008), approaches to noun and verb classification at the same time such as Pereira *et al.* (1993); Rooth *et al.* (1999), approaches to adjective classification such as Hatzivassiloglou and McKeown (1993); Bohnet *et al.* (2002); Boleda (2007), and approaches that apply across parts-of-speech, usually referred to as approaches to thesaurus induction such as Lin (1998a); McCarthy *et al.* (2003).

Semantic classifications are of great interest to computational linguistics, specifically regarding the pervasive problem of data sparseness in the processing of natural language, because classes of words can predict and refine properties of a word that received insufficient empirical evidence, with reference to words in the same class. For example, semantic verb classifications have up to now been used in applications such as word sense disambiguation (Dorr and Jones, 1996; Kohomban and Lee, 2005), machine translation (Dorr, 1997; Prescher *et al.*, 2000; Koehn and Hoang, 2007), document classification (Klavans and Kan, 1998), and also in psycholinguistic models of human sentence processing (Padó *et al.*, 2006). However, even though semantic classifications have already proven useful in many aspects, the large variety of semantic classifications also raises some questions.

(1) **Why there are so many classifications with the same target objects, why and how do they differ, and which of them is optimal?**

We find various classifications of the same part-of-speech even within the same language. For example, there are at least four different manual semantic classifications of German verbs, a process-based classification by Ballmer and Brennenstuhl (1986); *GermaNet* (the German version of WordNet), cf. Hamp and Feldweg (1997); Kunze (2000); the German version of FrameNet as compiled by the SALSA project, cf. Erk *et al.* (2003b); and the semantic classes by Schulte im Walde (2003, chapter 2). Obviously, the background of the authors of the classifications, their goals and their strategies directed the development of the verb classes. But even when two approaches classify verbs in a common language and according to a common framework, the results may still disagree. For example, Schulte im Walde defined semantic classes for German verbs by similar criteria as FrameNet; however, while Schulte im Walde classifies the *manner of motion (MOM)* verbs *eilen* and *hasten* (both meaning: 'to rush, to hurry') into a MOM subclass *rush*, FrameNet does not distinguish speed of motion into a separate class and groups these verbs with other *self motion* verbs. Both classes and assignments are plausible, but focus on different properties of the verbs –one concentrating on the rush, the other on an agent as mover. It seems that such differences are not fundamental flaws in the resources, but rather inherent in the task of semantic

classification. Schulte im Walde and Erk (2005) explored this intuition, by addressing the questions of *why there are so many classifications with the same target objects, why and how they differ, and whether any of them is optimal*, presenting a manual study on German verb classifications and the manner of motion domain that compared the classifications with respect to their motivation, class organisation, and sense and feature distinctions. The paper is included in this volume in Section 2.1, as one motivation for the subsequent application of human data. Summarising the results of the study, we found a small set of central sense distinctions that appear in all or almost all resources, and in addition there are idiosyncratic criteria that are used by few or only one resource. While the classifications often disagree, this is not a question of right or wrong but rather results from them focusing on different meaning criteria. These results are relevant from a computational perspective, addressing the automatic acquisition of semantic classes: The decision about which criteria are relevant for a semantic classification influences both the experiment setup (with regard to feature selection) and the choice of a manually constructed gold standard for evaluation. Knowing that each manual resource has its strengths and weaknesses, but that the resources nevertheless agree in central semantic dimensions, it is promising to combine several lexical resources, so their combination strengthens central meaning aspects while weakens marginal ones. Using only one individual resource, on the contrary, puts a bias on issues such as salient feature extraction, and evaluation scores, according to the dimensions of the respective resource.

(2) **Can human data support the identification of classification features and the evaluation of classification models?**

This question addresses a core issue of this work: whether and how we can exploit human data to automatically determine and evaluate theoretically adequate models of semantic classifications. An automatic acquisition of semantic classes relies on the definition of several parameters of the classification approach, most obviously the choice of relevant and representative objects to be classified, the properties of the objects used in the classification procedure, and the algorithm itself, for class formation and assignment. While we postpone the issue of feature selection for specific semantic classifications to Section 1.2 and the choice of the classification procedure to Section 1.3, the current section addresses the theoretical adequacy of feature selection in general, and the theoretical adequacy of classification results, as both issues refer to human data. The following two Subsections 1.1.1 and 1.1.2 explain our ideas in some detail.

1.1.1 Human Associations and Semantic Verb Classes

As mentioned before, various automatic methods have been applied to induce semantic classes from corpus data. Most such methods are borrowed from Artificial Intelligence which offers us a large variety of classification and clustering approaches.[4] Depending on the types of classes to

[4]The term *classification* comprises both *classification* and *clustering* approaches. The two classes of approaches are different with respect to pre-existing knowledge about the classes: In classification approaches, the desired

be induced, the techniques vary their choice of objects to be classified, and their classification algorithm. However, another central parameter for the automatic induction of semantic classes is the selection of the object features.

The feature selection should model the kinds of semantic relatedness between the words within the desired semantic classes. For example, Merlo and Stevenson (2001) classified 60 English verbs which alternate between an intransitive and a transitive usage, and assigned them to three verb classes according to the semantic role assignment in the frames; their verb features were chosen such that they modelled the syntactic frame alternation proportions and also heuristics for semantic role assignment. However, when it comes to larger-scale classifications with several hundreds of objects as investigated by e.g., Korhonen *et al.* (2003); Schulte im Walde (2006b); Joanis *et al.* (2008), who model verb classes by exploiting similarities at the syntax-semantics interface, it is not clear which features are the most salient. In these approaches, the verb features need to relate to a behavioural component (modelling the syntax-semantics interplay), but the set of features which potentially influence the behaviour is large, ranging from structural syntactic descriptions, prepositional phrases and argument role fillers to adverbial adjuncts. Even though there is agreement on the usefulness of some features, such as verb subcategorisation frames, there is still ongoing work on enlarging and optimising the set of features. An example of potentially relevant features that so far have only had an unsatisfying effect on semantic verb classes are selectional preferences (Schulte im Walde, 2000, 2006b). Even though, from a theoretical point of view, selectional preferences are potentially useful to characterise verb semantics, so far no computational approach has succeeded in implementing them such that they support the semantic class definitions. Furthermore, as mentioned above and illustrated by Schulte im Walde and Erk (2005), various approaches to manual semantic classifications might differ with respect to salient object properties.

To summarise, assuming that one is interested in a theoretically adequate choice of features to describe the objects to be classified semantically, what is missing is a general instrument to suggest and evaluate the semantic appropriateness of features. This work suggests human data as one such instrument, more specifically: *association norms*. Association norms –words that are called to mind by a set of stimulus words, as collected from human participants in association experiments– have a long tradition in psycholinguistic research, where they have been used for more than 30 years to investigate semantic memory, making use of the implicit notion that associates reflect aspects of word meaning (Tanenhaus *et al.*, 1979; McKoon and Ratcliff, 1992; Plaut, 1995; McRae and Boisvert, 1998, among others). Given that the meaning aspects of words are exactly what underlies any semantic classification, we take advantage of this long-standing notion: We exploit a collection of associations to check the salience of previously suggested features. Of course, we do not assume that there is an overall optimal set of features in automatic semantic classification. The goal is rather to determine (a) whether association norms represent an appropriate source of information for aspects of meaning and relatedness that are generally applicable to semantic classification, and (b) whether they in addition help to identify

classes are known in advance; in clustering approaches, the exploration of the classes and their structure is part of the task. Within the course of this work, we rely on classification vs. clustering approaches depending on the task.

resources and methods of lexical acquisition that improve the automatic induction of meaning aspects.

According to these overall goals, our contribution to this issue is included in Sections 2.2 and 2.3: On the one hand, ***association norms of German verbs and nouns are analysed*** in some detail, focusing on the question whether and how the various response types to the stimuli can be characterised by existing large-scale lexical and corpus resources. The underlying assumption is that semantic associates reflect highly salient linguistic and conceptual features of the stimulus word. Given this assumption, identifying the types of information provided by speakers and distinguishing and quantifying the relationships between stimulus and response can serve our goals as defined above.

Schulte im Walde *et al.* (2008b), incorporated as Section 2.2, provide a morpho-syntactic analysis, an analysis of syntax-semantic verb-noun functions, a co-occurrence analysis, and an analysis of semantic relations between stimuli and responses. We complemented each analysis with a discussion of the impact that it might have on NLP tasks and applications. A focus of the analyses is on issues related to word properties and word relations, i.e., addressing the task of modelling word meaning by empirical features in data-intensive lexical semantics, and providing insight into which types of semantic relations are treated as important by the speakers of the language, thus addressing two core issues within the automatic acquisition of semantic classes.

Schulte im Walde and Melinger (2008), incorporated as Section 2.3, provide a follow-up study on the co-occurrence analysis of the German verb association norms: We exploited the co-occurrence hypothesis (Miller, 1969; Spence and Owens, 1990), which holds that semantic association is related to the textual co-occurrence of the stimulus-response pairs. Within the article, we conducted a descriptive and in-depth examination of the distributional properties of the stimulus-associate pairs across context windows. In addition to replicating the basic experiments by Spence and Owens, we also broke the analysis down into various categories which had been independently identified as distributionally interesting (e.g., by Deese (1965); Clark (1971); McEvoy and Nelson (1982); Schulte im Walde *et al.* (2008b)), such as association strength, corpus frequency of the stimuli and responses, response part-of-speech, window direction, etc. Furthermore, we added analyses that question some of the intuitive conclusions from early work on the co-occurrence assumption, such as the association chain effect.

The common goal of all the analyses within the two above articles is not only to identify the characteristics of the words and the relations between words in the association norms, but at the same time –and crucial for the automatic acquisition of semantic classes– to identify resource- and corpus-based methods of how to extract word properties and word relations. Our results suggest co-occurrence information for an appropriate usage in empirical descriptions of word properties, an important insight since co-occurrence information is essentially less expensive (because no high-level pre-processing such as parsing is necessary), and therefore easier to obtain –especially in languages with few NLP resources available– than annotated data. A further contribution concerns the generalisation of a specific choice of features: Distributional feature descriptions as well as semantic relationships only cover the "average" of word meaning aspects. However, if

one is concerned with specifying word properties and word-word relations with respect to individual words, the semantic class and the frequency range of that word should be taken into account

Following the work on analysing association norms, the second part of the contribution here explored ***how to utilise association norms within classification approaches***. Schulte im Walde (2008b), incorporated as Section 2.4, investigated whether the association norms for German verbs as mentioned above can help us to identify salient features for semantic verb classification. In a first step –to explore the structure of a classification based on associations, and the assignment of individual verbs– I applied a cluster analysis to German verbs, based on their associations, and validated the resulting verb classes against standard approaches to semantic verb classes. Then, I performed various clusterings on the same verbs using standard corpus-based feature types, and evaluated them against the association-based clustering as well as GermaNet and FrameNet classes. Comparing the cluster analyses provided an insight into the usefulness of standard feature types in verb clustering, and assessed shallow vs. deep syntactic features, and the role of corpus frequency. Summarising the results, the article showed that (a) there is no significant preference for using a specific syntactic relationship (such as direct objects) as nominal features in clustering; (b) that simple window co-occurrence features are not significantly worse (and in some cases even better) than selected grammar-based functions; and (c) that a restricted feature choice disregarding high- and low-frequency features is sufficient. Finally, by applying the feature choices to GermaNet and FrameNet verbs and classes, I addressed the question of whether the same types of features are salient for different types of semantic verb classes. The variation of the gold standard classifications demonstrated that the clustering results are significantly different, even when relying on the same features. In a further article (Melinger *et al.*, 2006), incorporated as Section 2.5, we applied a soft-clustering approach to the association norms for German nouns that have been mentioned above. We showed that –based on the associations– the resulting cluster analysis could be applied to predict noun ambiguity and to discriminate the various senses of polysemous target nouns. Both articles illustrated that association norms are not only a useful source of information for aspects of meaning that are generally applicable to semantic classification, but also that they can be used in combination with computational methods of lexical acquisition, to improve the automatic induction of meaning aspects.

1.1.2 Human Judgements and Semantic Adjective Classes

A second issue that exploits human data with respect to the automatic acquisition of semantic classes is concerned with the evaluation of semantic classifications. Basically, there are three ways to evaluate a computational model: (i) by human assessment, (ii) by comparison against a gold standard, and (iii) by incorporating the model into an application. Concerning (ii), ***evaluation by comparison against a gold standard***, I demonstrated in Schulte im Walde (2003, chapter 4) that there is no absolute scheme for evaluating the result of a cluster analysis. A variety of evaluation measures from diverse areas such as theoretical statistics (Rand, 1971; Fowlkes and Mallows, 1983; Hubert and Arabie, 1985), web-page clustering (Strehl *et al.*, 2000) and corefer-

ence resolution (Vilain *et al.*, 1995) do exist, but a priori it is not clear which one fits best to the linguistic task of inducing semantic classes. Nevertheless, various ways of evaluating a semantic classification against a gold standard have been suggested. For example, Hatzivassiloglou and McKeown (1993) defined a recall and precision measure for evaluating a cluster analysis of adjectives on the basis of the common cluster membership of object pairs in the clustering and the gold standard. Schulte im Walde and Brew (2002) presented an adjusted version of their measures that explicitly takes linguistic constraints into account. This adjusted measure was subsequently applied also by Korhonen *et al.* (2003). In addition, Stevenson and Joanis (2003) and Korhonen *et al.* (2003) applied *accuracy*, which requires a reference from each induced cluster to a gold standard class, according to the majority of member overlap, and calculates the proportion of verbs correctly classified, similarly to the individual entity links in Bagga and Baldwin (1998) and Luo (2005). Furthermore, Stevenson and Joanis (2003) exploited the *mean silhouette* (Kaufman and Rousseeuw, 1990) which determines how well the resulting clusters separate the objects according to the underlying data, and they adopted the *adjusted Rand index* from my analysis in Schulte im Walde (2003, chapter 4). Summarising, over the years a number of evaluation measures for semantic classes against a gold standard have been defined. Two basic problems remain, though: First, especially for large-scale semantic data there is not necessarily a gold standard available, and furthermore –as exemplified by Schulte im Walde (2006c)– if there are several gold standards, the evaluation results might differ substantially. And second, even if there is a unique gold standard to use, there is a variety of evaluation measures whose theoretical adequacy is not obvious, and the results of applying several measures might differ in their ranking. Concerning (iii), ***incorporating the model into an application***, there are various NLP systems that incorporated semantic classes, as mentioned earlier. The focus of such systems is, however, not on the theoretical adequacy of the classifications, but on their successful application, and these two motivations are not necessarily correlated. At least, it does not follow from a successful application of a semantic classification that the classification is linguistically sound.

This work thus focuses on evaluation strategy (i), ***human assessment***, as relying on human data seems most promising with respect to the theoretical adequacy of the classification model. Experiments that gather human judgements on linguistic phenomena are, however, very difficult to design for two main reasons. First, the agreement between annotators decreases with the complexity of the task (Artstein and Poesio, 2008). Second, in order to obtain judgements on a large scale, the experiments need to address non-expert participants in addition to expert participants, but it is deemed to cause difficulties for the non-expert judges if linguistic background is required. Boleda *et al.* (2008), incorporated in this work as Section 2.6, report on a large-scale experiment for gathering human judgements with respect to a semantic classification of Catalan adjectives. The goal of our experiment was to classify 210 Catalan adjectives into three semantic classes. The experiment was directed at non-expert native speakers and administered over the Internet, collecting data from 322 participants. We assessed the degree of inter-annotator agreement through an innovative methodology based on observed agreement and kappa, and used weighted versions of these measures to account for partial agreement in polysemous assignments. We then performed a series of post-hoc analyses on the human judgements to distinguish disagreement caused by the task as opposed to that caused by the experimental design, thus pointing to spe-

cific difficulties in both aspects of the research. The methodology developed in this article might therefore prove useful for the design of experiments for related tasks.

1.2 Feature Exploration for Specific Semantic Classifications

Chapter 3 addresses the theoretical adequacy of feature selection for an automatic semantic classification. All three articles within this chapter start out with word properties derived from the theoretical literature, suggest ways to induce the respective features from corpus data, and assess the effect of the feature selections on word relatedness and semantic classification. In general, the effects of the feature choices on the semantic models correspond to what theory predicts. Our explorations go beyond this correspondence, though, by suggesting and assessing the respective computational models and pointing to automatic means for further work.

The contribution to this line of research is split into two parts: Sections 3.1 and 3.2 focus on features at the syntax-semantics interface that are relevant to the description and the semantic relatedness of German particle verbs, exploring the specificities of this subclass of verbs. Section 3.3 focuses on the contribution of morphological vs. syntactic features with respect to a semantic classification of Catalan adjectives. In the following, the articles and their contributions are described in some more detail.

The article in Section 3.1 lays the foundation of the quantitative work on German particle verbs. German particle verbs are productive compositions of a base verb and a prefix particle, whose part-of-speech varies between open-class nouns, adjectives, and verbs, and closed-class prepositions and adverbs. My work concentrates on prepositional particle verbs, such as *abholen*, *anfangen*, *einführen*. Particle verb senses are situated on a continuum between transparent (i.e., compositional) or opaque (i.e., non-compositional) meaning, with respect to their base verbs. For example, *abholen* ‘fetch’ is transparent with respect to its base verb *holen* ‘fetch’, *anfangen* ‘begin’ is opaque with respect to *fangen* ‘catch’, and *einsetzen* has both transparent (e.g., ‘insert’) and opaque (e.g., ‘begin’) verb senses with respect to *setzen* ‘put/sit (down)’. A specificity of German particle verbs is that they may change the syntactic behaviour of their base verbs: the particle can saturate or add an argument to the base verb’s argument structure, cf. example (1.1) from Lüdeling (2001). Theoretical investigations (Stiebels, 1996) and corpus-based work (Aldinger, 2004) have demonstrated that those changes are quite regular.

(1.1) Sie *lächelt*.
‘She smiles.’
*Sie *lächelt* [$_{NP_{acc}}$ ihre Mutter].
‘Sie smiles her mother.’
Sie *lächelt* [$_{NP_{acc}}$ ihre Mutter] *an*.
‘Sie smiles her mother at.’

Even though German particle verbs constitute a significant part of the verb lexicon, most work so far has been devoted to theoretical investigations. As far as I know, no other work has provided any quantitative analysis of German particle verbs, except for a corpus-based analysis by Aldinger (2004) that extracted alternation patterns for subcategorisation frames of particle and base verbs.

My first article addressed three basic issues concerning German particle verbs: (1) It describes how the particle verbs were identified with the help of a statistical parser for German (which is a difficult task per se, as particle and base verb are obligatorily adjacent and morphologically combined in verb-final sentences, but separated in verb-second and verb-first sentences). (2) It extracted and compared subcategorisation frame types and their argument fillers for particle verbs vs. their base verbs. The results corresponded to the theoretical definitions, illustrating that transparent as well as opaque particle verb senses might or might not undergo a change with respect to their base verbs; but that concrete nominal argument fillers indicatee (dis)agreement of verb senses and thus were pointers to the degree of transparency. (3) A simple standard approach to distributional similarity used the distributional features in (2) to predict the semantic relatedness (and thus, the degree of transparency) between particle and base verbs.

The article in Section 3.2 continues on the work in the previous article. It provides a nearest neighbour analysis for German particle verbs as preliminary work towards a semantic classification: I.e., it identified German verbs that are semantically most related to German particle verbs, based on various standard features at the syntax-semantics interface and a standard approach to measuring distributional similarity. The results illustrated that the syntax-semantics mapping hypothesis (that to a certain extent, the lexical meaning of a verb determines its behaviour, particularly with respect to the choice of its arguments, cf. Pinker (1989); Levin (1993)) does not apply to particle verbs as it does to verbs in general: Transparent particle verb senses are semantically related to their base verbs, but nevertheless do not necessarily agree with them in their syntactic behaviour. And since we know that semantically related non-prefixed verbs show agreement in their behaviour to a large extent, we assume that the frame mismatch transfers from the base verbs to other verbs in their respective semantic class. This means that a syntactic description of transparent particle verbs and semantically related verbs is not expected to show strong overlap. For opaque particle verb senses, it is more difficult to make strong statements. Since they compositionally represent idioms, I argued that they undergo the syntax-semantic relationship, i.e., that they behave similarly as semantically related verbs. The most successful features were nominal preferences (i.e., nominal heads of arguments), either with or without reference to the respective subcategorisation frames. The comparison of the various feature distributions thus demonstrated that –in accordance with theoretical expectations– syntactic descriptions are not much help in defining the semantics of German particle verbs (because of the syntactic change in argument structure) but that the nominal arguments are useful indicators of particle–base verb relatedness.

In addition to the feature exploration for German particle verbs, the article provides two more contributions to the issues of this work: First, it used various gold standards to assess the nearest neighbours of the particle verbs (GermaNet, a dictionary of synonyms and antonyms (Bulitta and Bulitta, 2003), and our collection of semantic associates to verbs, cf. Section 2.2). On the one

hand, the article then showed that the precision of the nearest neighbours varied strongly with the gold standard (thus indicating that there are severe differences in gold standard resources, even if they have a common goal, cf. Section 1.1.2). On the other hand, it confirmed –with the associations representing the most successful gold standard, according to the precision values– that associations are indeed a useful human data source with respect to semantic word relatedness. A second further contribution applied Latent Semantic Analysis (LSA), cf. Deerwester *et al.* (1990), to the feature distributions, and then identified the nearest neighbours on the basis of the LSA matrix. The result was that for the task of identifying semantic nearest neighbours on the basis of specific verb-noun relations, the task precision suffered from reducing the matrix information by LSA. Only when using the original frequencies and with certain dimensionality, the task-relevant information was preserved. However, for the purpose of time-saving experiments, a single specific reduction was sufficient. In conclusion, it is advisable to apply LSA (and invest the time to find the optimal dimensions) only in cases where succeeding experiments profit from the reduced number of features.

The final article of Chapter 3 in Section 3.3 explored features for Catalan adjectives. So far, there is no unique established standard of semantic classes for Catalan adjectives, and furthermore there are various suggestions towards salient features of such classifications, cf. Boleda (2007, chapter 3). Following the ontological framework by Raskin and Nirenburg (1998), we suggested three semantic classes of Catalan adjectives, and investigated morphological and syntactic features within a decision tree approach, to check on their reliability as semantic properties. In Catalan, there is an obvious relationship between the derivational type of an adjective and its semantic class. Therefore, the simplest classification strategy was to associate each derivational type with a semantic class. And indeed, the morphological features succeeded with high accuracy, but they failed in cases of non-compositional meanings. Syntactic features (unigrams, bigrams, and the syntactic functions of the adjectives) were slightly more successful than the morphological features, but systematically confused two of the semantic classes which were syntactically not sufficiently distinct. A combination of the two types of features outperformed both individual sets. The article not only showed the usefulness of the various features in automatic semantic classification, but also shed light on the characteristics of each class, thus contributing to their theoretical profiles.

1.3 Approaches to Modelling Polysemy in Semantic Classification

Even though polysemy is a pervasive phenomenon in semantic classification, few models have explicitly included polysemy within their automatic classification approaches. For example, concerning semantic verb classifications, to our knowledge so far only Pereira *et al.* (1993); Rooth *et al.* (1999); Korhonen *et al.* (2003) suggested models that deal with polysemous verbs; concerning semantic adjective classification, no automatic approach has previously dealt with polysemous adjectives. In my opinion, there are two main reasons for this lack of approaches

considering polysemy: (1) In addition to the fact that there is still ongoing work on enlarging and optimising the feature sets used in automatic semantic classification (as mentioned before), it is even more difficult to determine a theoretically adequate model for semantic classification *with* than *without* incorporating polysemy. Such models are mathematically more complex and in addition raise further questions, e.g., how many classes (per object) are appropriate when polysemy is considered? And (2), it is even less clear how to evaluate a semantic classification that allows multiple class-per-object assignments than in the simpler, monosemous case, cf. Section 1.1.2. The most straightforward evaluation of such a model is its integration into an application; but it does not necessarily follow from a successful application of a semantic classification that the classification is linguistically sound, as mentioned before.

This work incorporates two approaches to modelling polysemy in semantic classes. The first approach, incorporated as Section 4.1, classifies Catalan adjectives, relying on multi-label classification and an ensemble classifier; the second approach, incorporated as Section 4.2, classifies English verbs (and is applicable to verb classification in all languages for which the WordNet functions provided by Princeton University are available), using a complex approach that combines an EM classifier and the MDL principle.

Boleda *et al.* (2007) suggested multi-label classification with respect to the three Catalan adjective classes mentioned already. Adjective classification was performed within a two-level architecture for multi-label classification: first, make a binary decision on each of the classes, and then combine the classifications to achieve a final, multi-label classification. We therefore decomposed the global decision on the (possibly polysemous) class of an adjective into three binary decisions: *Is it class*$_1$ *or not? Is it class*$_2$ *or not? Is it class*$_3$ *or not?* The individual decisions were then combined into an overall classification that included polysemy. For example, if a lemma was classified both as class$_1$ and as class$_2$ in each of the binary decisions, it was deemed polysemous (class$_{1+2}$). The motivation behind this approach was that polysemous adjectives should exhibit properties of all the classes involved. As a result, positive decisions on each binary classification could be viewed as implicit polysemous assignments. The classification architecture is very popular in Machine Learning for multi-label problems, cf. Schapire and Singer (2000); Ghamrawi and McCallum (2005), and has also been applied to NLP problems such as entity extraction and noun-phrase chunking (McDonald *et al.*, 2005). As classifier for the binary decisions we chose decision trees; as feature descriptions, we used morphological, syntactic and semantic indicators, an extension of those in Boleda *et al.* (2005). A comparison of the individual binary decisions with the combined decisions based on the multi-label classifier showed that the accuracy of the multi-label classification outperformed the accuracy of the individual decisions; furthermore, the differences between the various feature sets are much clearer in the combined vs. the individual decisions. We concluded that polysemy acquisition naturally suits multi-label classification architectures.

In a further part of the article, we implemented an ensemble classifier, a type of classifier that has received much attention in the Machine Learning community in the last decade (Dietterich, 2002). When building an ensemble classifier, several class proposals for each item are obtained, and one of them is chosen on the basis of majority voting, weighted voting, or more sophisticated

decision methods. It has been shown that in most cases, the accuracy of the ensemble classifier is higher than the best individual classifier (Freund and Schapire, 1996; Dietterich, 2000; Breiman, 2001). Within NLP, ensemble classifiers have been applied, for instance, to genus term disambiguation in machine-readable dictionaries (Rigau *et al.*, 1997), using a majority voting scheme upon several heuristics, and to part-of-speech tagging, by combining the class predictions of different algorithms (van Halteren *et al.*, 1998). The main reason for the general success of ensemble classifiers is that they gloss over the biases introduced by the individual systems. Our implementation used the different levels of description as different subsets of features, and applied majority voting across the class proposals from each level. The ensemble classifier boosted the performance of the system beyond the best single type of information and is thus a more adequate way to combine the linguistic levels of description than simply merging all features for classification.

Schulte im Walde *et al.* (2008a) presented an innovative, complex approach to semantic verb classification that relied on selectional preferences as verb properties. The underlying linguistic assumption for this verb class model was that verbs which agree on their selectional preferences belong to a common semantic class. The model was implemented as a soft-clustering approach, in order to capture the polysemy of the verbs. The training procedure used the Expectation-Maximisation (EM) algorithm (Baum, 1972) to iteratively improve the probabilistic parameters of the model, and applied an instantiation of the Minimum Description Length (MDL) principle (Rissanen, 1978) by Li and Abe (1998) to induce WordNet-based selectional preferences for arguments within subcategorisation frames. As result, the model provided soft clusters with two dimensions (verb senses and subcategorisation frames with selectional preferences). The model is generally applicable to all languages for which WordNet exists, and for which the WordNet functions provided by Princeton University are available. For the purposes of the paper, we chose English as a case study.

Our model is an extension of the latent semantic clustering (LSC) model (Rooth *et al.*, 1999) for verb-argument pairs. In comparison to our model, the LSC model only considers a single argument (such as direct objects), or a fixed number of arguments from one particular subcategorisation frame, whereas our model defines a probability distribution over all subcategorisation frames. Furthermore, our model specifies selectional preferences in terms of general WordNet concepts rather than sets of individual words. In a similar vein, our model is both similar and distinct in comparison to the soft clustering approaches by Pereira *et al.* (1993) and Korhonen *et al.* (2003). Pereira *et al.* (1993) suggested deterministic annealing to cluster verb-argument pairs into classes of verbs and nouns. On the one hand, their model is asymmetric, thus not giving the same interpretation power to verbs and arguments; on the other hand, the model provides a more fine-grained clustering for nouns, in the form of an additional hierarchical structure of the noun clusters. Korhonen *et al.* (2003) used verb-frame pairs (instead of verb-argument pairs) to cluster verbs relying on the Information Bottleneck (Tishby *et al.*, 1999). They had a focus on the interpretation of verbal polysemy as represented by the soft clusters. The main difference of our model in comparison to the above two models is, again, that we incorporated selectional preferences (rather than individual words, or subcategorisation frames).

Within the scope of the article, the model was assessed by a language-model-based evaluation. The evaluation showed that after 10 training iterations the verb class model results were above the baseline results. Our model is potentially useful for lexical induction (e.g., verb senses, subcategorisation and selectional preferences, collocations, and verb alternations), and we plan to exploit the model with respect to its theoretical adequacy in this vein in future work.

1.4 Summary

The introduction to this work has brought together a selection of recent publications that all address a common topic: how to bridge the gap between the theoretical adequacy, the acquisition potential, and the successful application of a model of semantic word classes. A focus of this work is on modelling word properties and definitions of relatedness between words that are crucial within the automatic acquisition of semantic classes, addressing characteristics of semantic classes as well as classification approaches that support the automatic acquisition of such classes.

Following an overview of the contributions of this work, Sections 1.1 to 1.3 provided motivations and some details of the three main parts: (i) As a major source for the theoretical adequacy of the semantic knowledge, we exploit human data; association norms as a general instrument to suggest, apply and assess the semantic appropriateness of features in semantic classification, and human judgements to evaluate classification models. (ii) We explore empirical word properties for specific instantiations of semantic classes for which so far no unique established standard of semantic classes exist, German particle verbs and Catalan adjectives. Based on word properties derived from the theoretical literature, the correspondence between what theory predicts and the effect of the respective feature choice on the semantic models is explored. (iii) Addressing the lack of semantic classification models that so far have explicitly included polysemy into the automatic procedure, two approaches to modelling polysemy in semantic classes are suggested, multi-label classification, an ensemble classifier, and an Expectation-Maximisation classifier incorporating the Minimum Description Length principle. In the remainder of this work, the three main chapters are organised according to these three main parts.

Acknowledgements

Many thanks to Gemma Boleda, Aoife Cahill, Arndt Riester and Heike Zinsmeister for their valuable comments on previous versions of this introduction.